Athens Living Lab Ideathon

Edge Computing

Institute of Communication and Computer Systems (**ICCS**)

10 October 2022

Dr. Konstantinos V. Katsaros

Head of Intelligent Networks & Services
ISENSE Group / ICCS

Outline

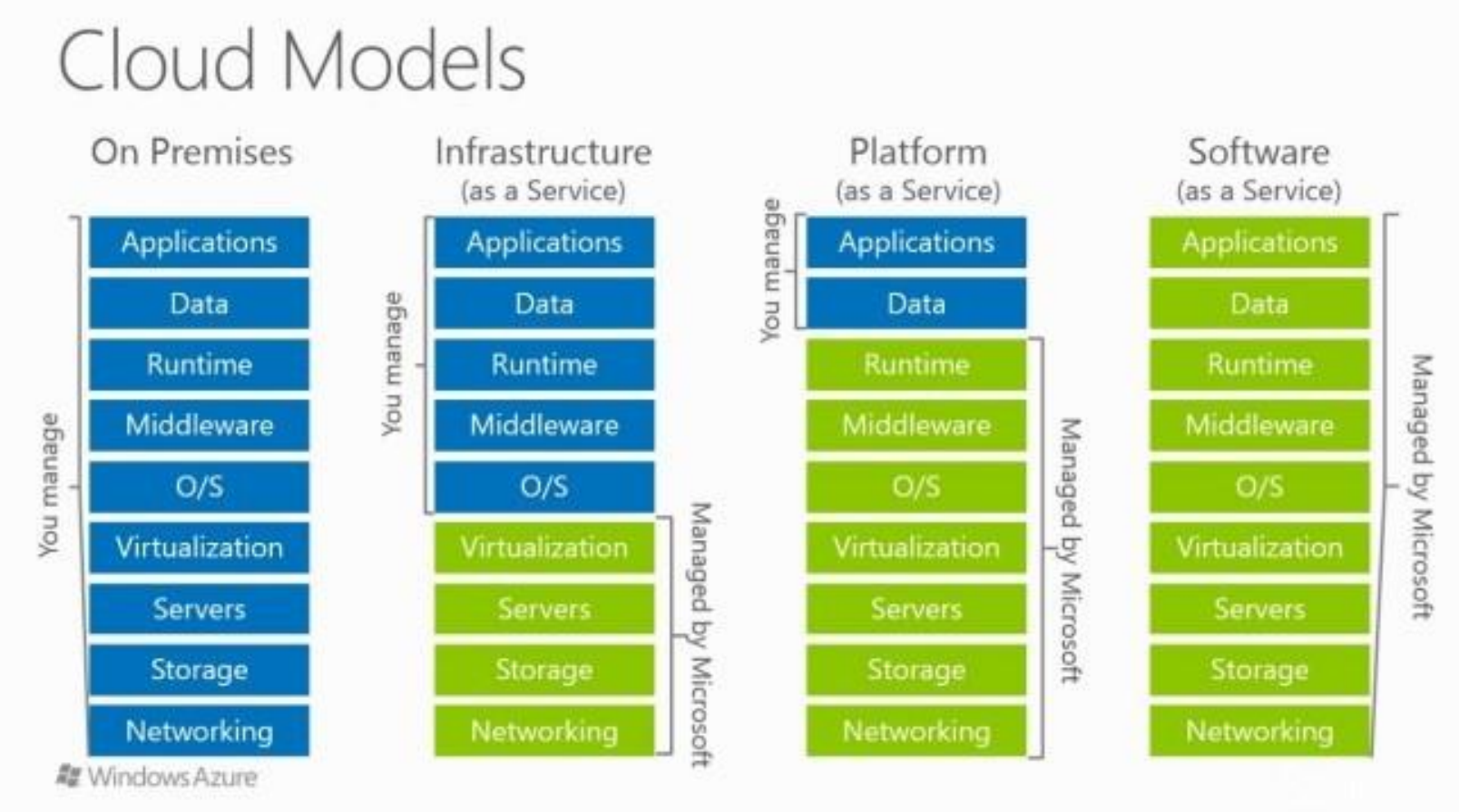
- Edge Computing in a nutshell
- Where is the Edge?
- Use Cases
- A Functional Perspective
  - Overview of Edge Computing internals
  - Developing and Deploying Edge Computing Applications
- Key Technologies
  - Existing Tools and Frameworks
- Challenges

Disclaimer
Edge computing is a very large area, impossible to sufficiently cover here!

5GLOGINNOV

The Basics: Cloud Computing

- Economies of scale
- Elasticity



Edge Computing in a Nutshell

Where is that exactly?

Concept

- Enable cloud computing capabilities at the edge of the network
- Vertical oriented
- Close integration with Radio Access Network (RAN)

Benefits

Why not earlier?

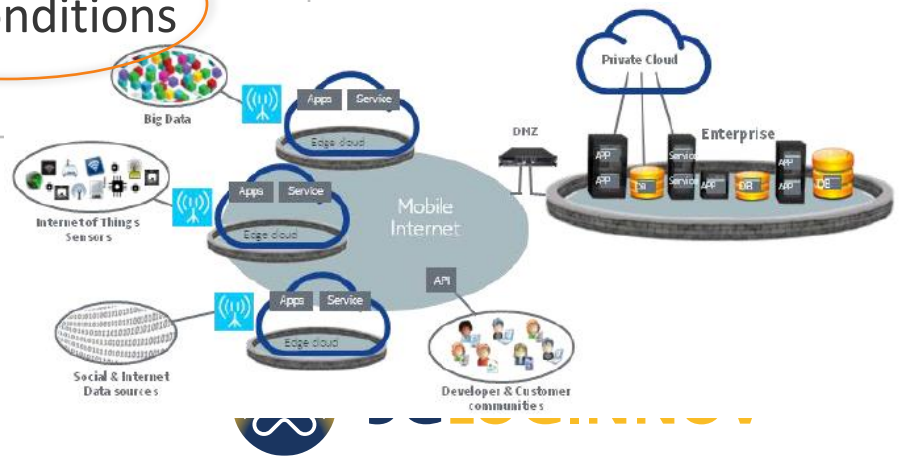
What else?

- Reduced latency
- Reduced network traffic
- Service optimization through **context-awareness** e.g., radio conditions
- …

Example application domains

- CCAM
- Smart Agriculture
- Manufacturing
- Content distribution

Specifically?



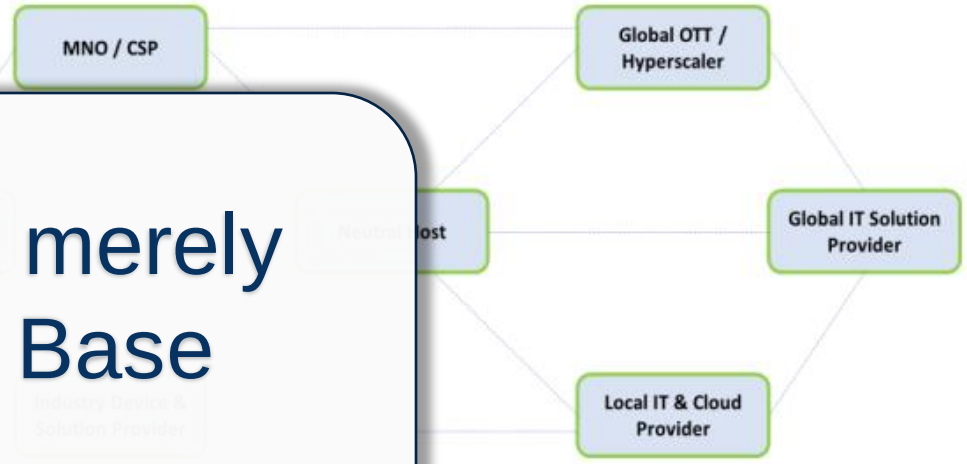
k.katsaros@iccs.gr

Where is the Edge?

Key players in Edge Computing

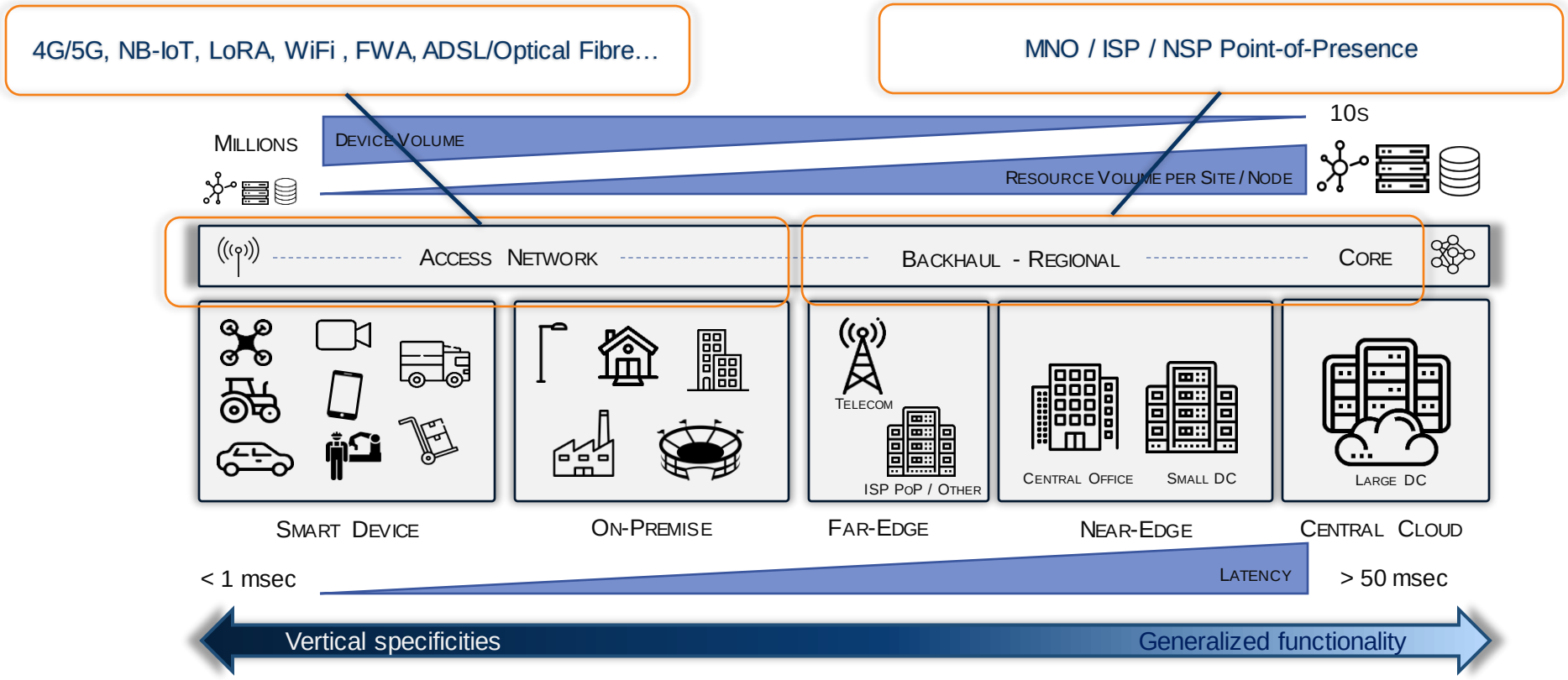
- Telecom/Mobile Network Operators (MNOs)
- Communication Service Providers (CSPs)
- Global OTT / Hyperscalers e.g., Google, Amazon
- Local IT and Cloud Providers
- Global IT Solution providers e.g., IBM, Oracle, HPE, etc.
- Telco vendors
  - Network Equipment Providers (NEPs) / Manage Service Providers (MSP)
- Global industry device & solution providers e.g. Siemens, ABB, etc.
- Neutral Host (provider) e.g., Barcelona City nodes

Edge Computing is **not** merely Servers connected to Base Stations.



Source: 5G-PPP White Paper on Edge Computing for 5G Networks

Where is the Edge?



The Compute Continuum

Use Cases

Use Cases

Automotive: Collective Environment Perception

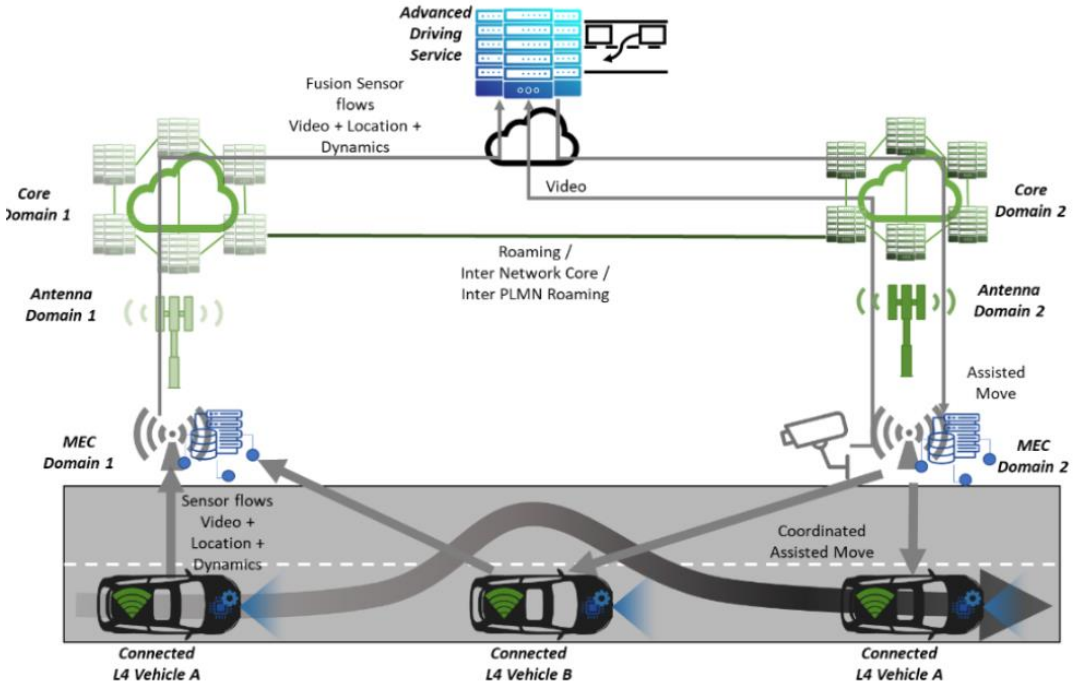
Edge Computing functionality / features

- Real-time exchange of vehicle sensor information
- Perception beyond local sensor range
- Aggregation, fusion, delivery of information

Applications

- Collision avoidance
- Automated manoeuvres

  e.g., overtaking and lane changing
- HD Maps



Collective perception environment based on Edge Computing (5G-MOBIX)

Use Cases

Manufacturing

Edge Computing functionality / features

- Real-time collection of component sensor data
- Real-time video analytics
- Closing the control loop: decision making & actuation
- Privacy/Security: non-public deployments

Applications

- Factory automation e.g., robotics w/ computer vision
- HMIs – AR/VR – Digital Twins
- HD Maps

5GLOGINNOV

Use Cases

Precision Agriculture

Edge Computing functionality / features

- Real-time video analytics
- Closing the control loop: decision making & actuation
- AR interfaces

Applications

- Precision Spraying
- Remote Disease Diagnosis
- Drone Based Monitoring

5GLOGINNOV

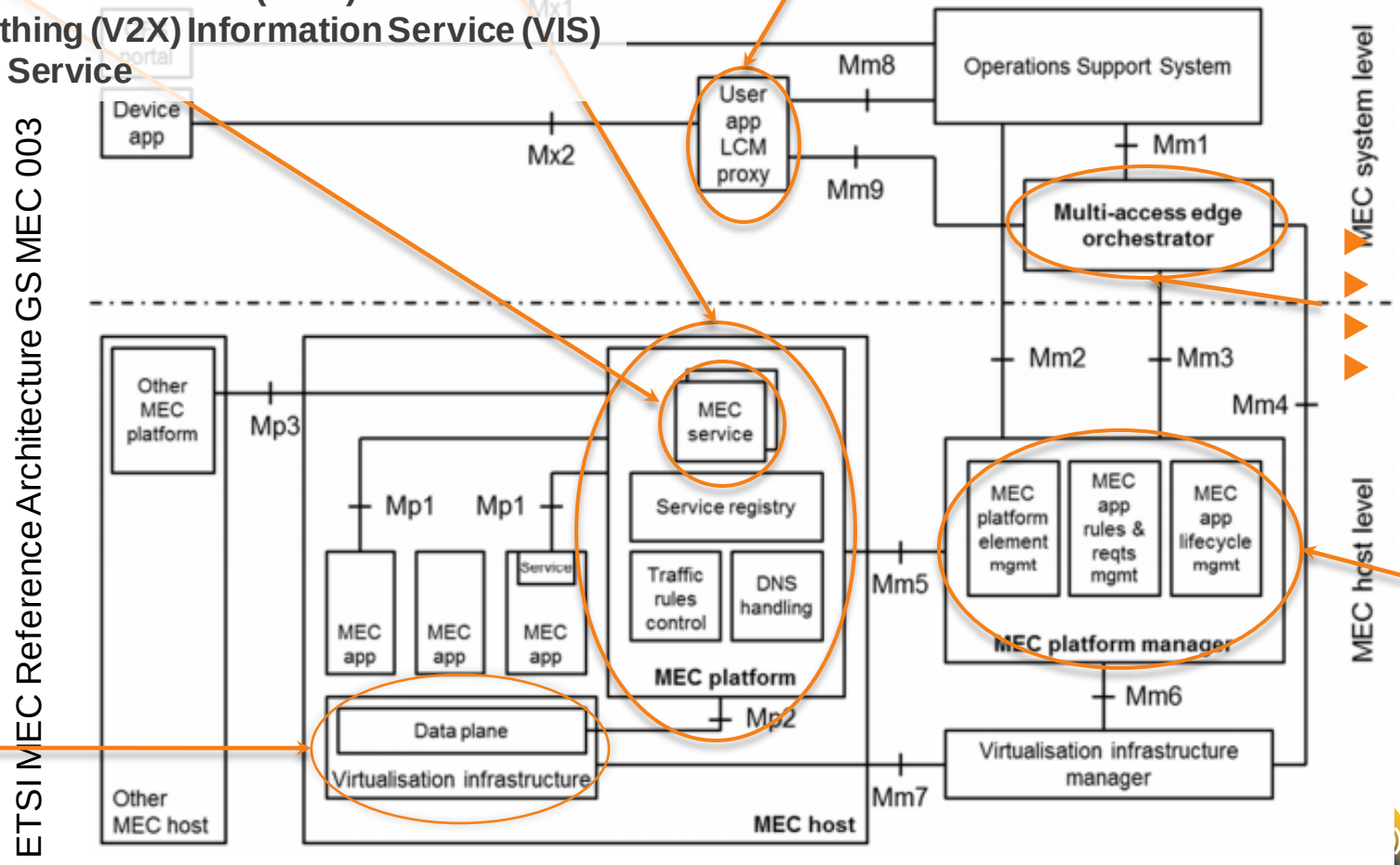
A Functional Perspective

Overview of Edge Computing internals

- Radio Network Information Service
- Location Service
- Traffic Management Service
- Fixed Access Information Service (FAIS)
- Vehicular-to-Everything (V2X) Information Service (VIS)
- WLAN Information Service

- App discovery /adv./consumption
- ME Services
- DNS & Traffic Rules

- Interface to UEs for:
  - On-boarding
  - Instantiation
  - Termination



- ME system overview
- App. Pckg. On-boarding
- ME Host Selection
- Triggering app. lifecycle events

- IaaS
- PaaS
- SaaS

- App. Lifecycle mgmt.
- EM functionality
- DNS & Traffic Rules mgmt.

5GLOGINNOV

Context-awareness: ETSI MEC Information Services

Radio Network Information Service	Location Information Service	Fixed Access Information Service (FAIS)	Vehicular-to-Everything (V2X) Information Service (VIS)	WLAN Information Service
• Radio Access Bearer • PLMN e.g., cell changes • L2 Measurements • …	• Specific UE location • Area / AP UE locations • Distance between Ues and/or specific point • Filtering of the above • …	• Device • Cable line • Optical network • …	• List of authorized PC5 UEs • Communication with other ISs • Predictive QoS notifications • …	• List of APs • WLAN capabilities • BSS Load • Station Data Rates • …

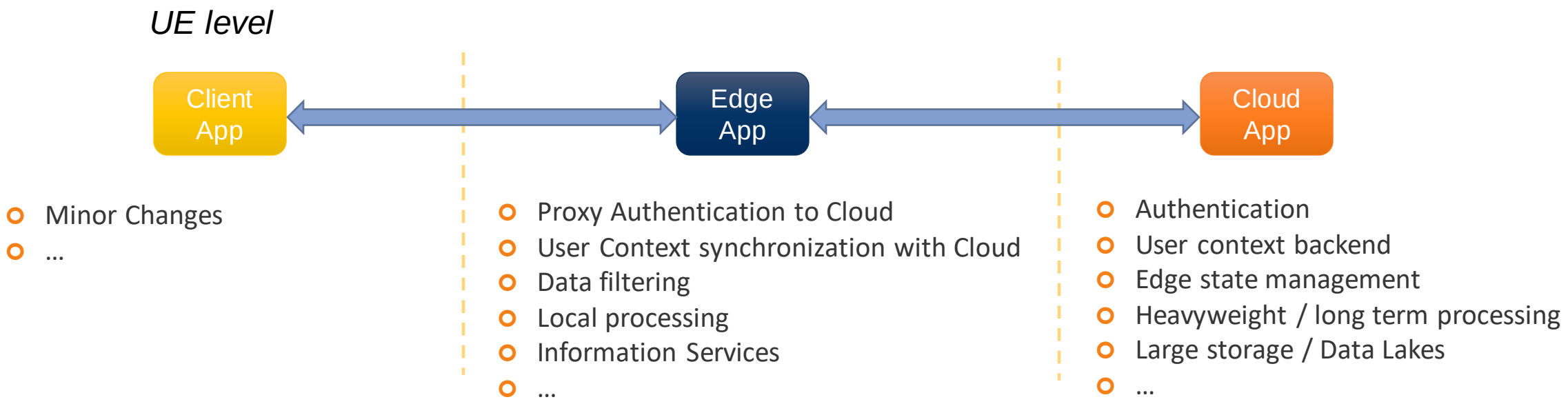
(*) This is a subset of the overall services

Offered by an ETSI MEC Platform, but other Services can be offered by Third Parties as well

5GLOGINNOV

Developing Applications for the Edge

- A paradigm change: Functional split
- Fit with: Microservices - Serverless architectures – Service Meshes



Functional split for single 3-tier environment: UE, Edge Node, Cloud
See also: ETSI White Paper No. 20 Developing Software for Multi-Access Edge Computing, February 2019

5GLOGINNOV

Deploying Applications at the Edge

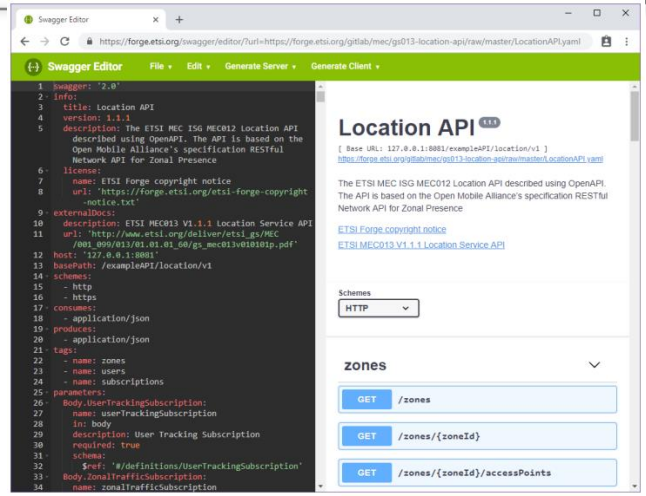
1. Packaging and on-boarding
- Prepare and sign VM/container
- Deliver (upload) to OSS →… → VIM e.g., Kubernetes
- Traffic/DNS rules, Use of Information Services

2. Instantiation & Operation
- UE or developer triggered
- Platform issues LCM requests
- Traffic/DNS, Information services configuration

3. UE – Edge communication
- Direct IP or DNS resolution
- Optional Edge App Mgmt functionality

4. Usage of Edge Services
- Information Services or 3rd party services
- RESTfull APIs

Further practical information:
- ETSI White Paper No. 20 Developing Software for Multi-Access Edge Computing, February 2019
- ETSI MEC Sandbox



5GLOGINNOV

Key Technologies

Key Technologies
Overview

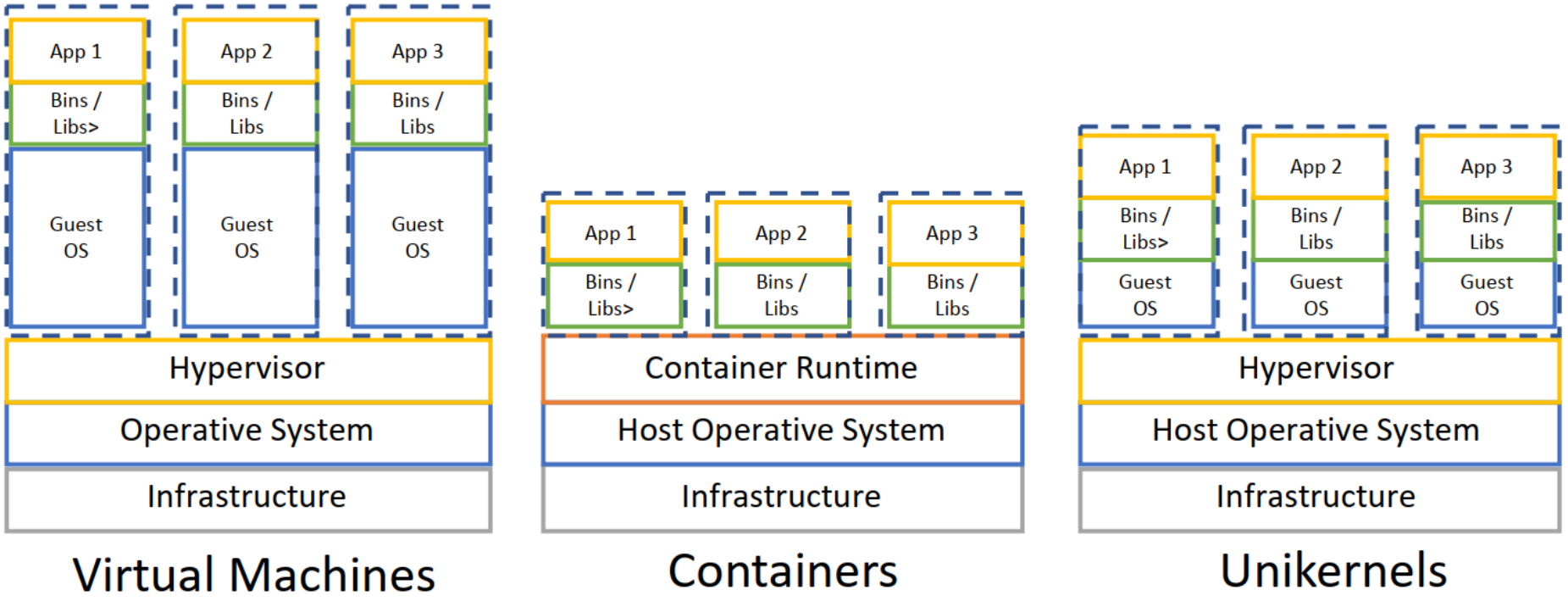
Resource Virtualization	Orchestration	Network Programmability	Acceleration
• Virtual Machines • Containers • Lightweight Virtualization	• Kubernetes • OSM • ONAP • *Other*	• SDN for Edge Computing • Data plane Programmability	• FPGAs • GPUs

Source: **5G-PPP White Paper on Edge Computing for 5G Networks**

5GLOGINNOV

Key Technologies

Resource Virtualization



Key Technologies

Orchestration

Generic Orchestration

- VM/Container Life-Cycle Management (LCM)
  - Instantiation, configuration, auto-scaling / load balancing, …
- Network Service LCM
  - Connectivity, traffic management, …

Edge Specificities

- Resource limitations
  - Platform footprint
  - Job prioritization
- Multi-node operation
  - State management, telemetry
- Mobility management
  - Session handover support: state-full / state-less applications
- Information services
  - Data collection/aggregation and exposure
- Service discovery & LCM implications

5GLOGINNOV

Key Technologies

Orchestration: Platforms & Tools

ONF 4G/5G Edge Platform; Focus on **connectivity/slicing**; Integration with SD-RAN / SD-CORE

Experimental; Lightweight; focuses on **Telco environments** 4G/5G; **ETSI MEC compliant; Integration w/ FlexRAN/OAI**

Micro-services based platform, focusing on **Industrial IoT- Interoperability**

Blueprints for 5G, AI, Edge IaaS/PaaS, IoT

Decentralized, generic platform for the edge. Pub/sub State Mgmt; Resource Abstractions

Telco Edge **Marketplace**; **Aggregates enterprise and operator infrastructure** on a global scale, harmonizing use

Experimental; Lightweight; focuses on **Telco environments** 4G/5G; **ETSI MEC compliant**

Powerful MANO tool targeting **ETSI NFV, Telco environments** e.g., 5G networks; focus on e2e services, slicing; works with OpenStack & Kubernetes

OpenSource distro of OpenShift (RedHat). Kubernetes distro; **Focus on DevOps and Multi-tenancy**

Google Application Mgmt Platform for distributed environments; Builds on Kubernetes.

Lightweight version (distro) of **Kubernetes** (different deployment architecture)

Lightweight version (distro) of **Kubernetes**

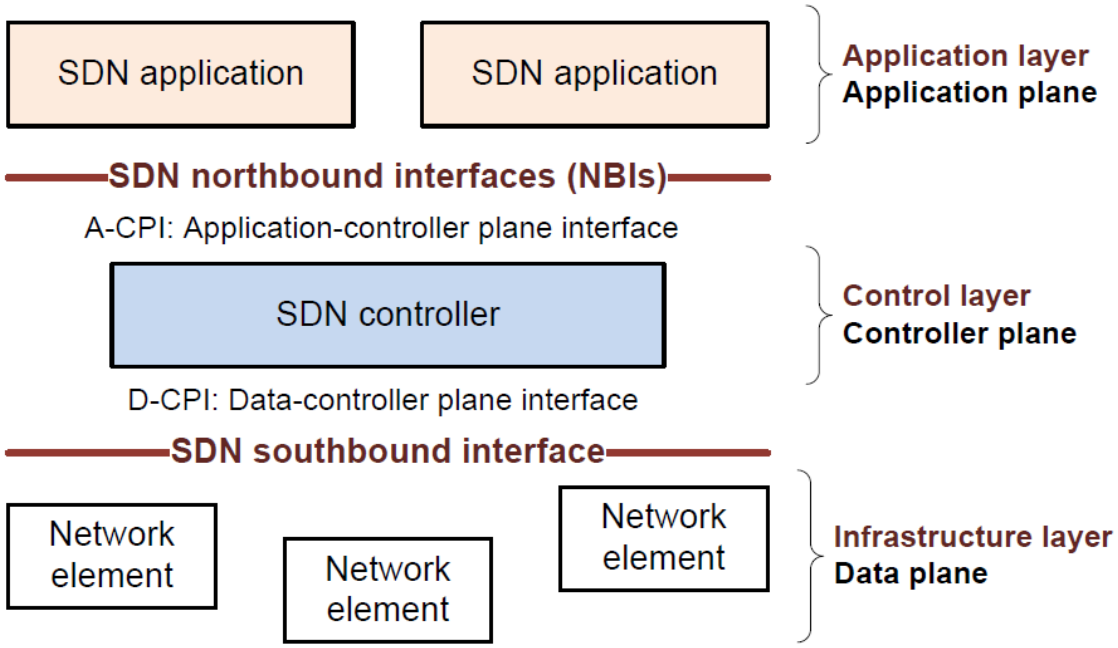
Powerful **cluster management** platform; Forms the basis for multiple Edge platforms; Can be deployed on micro-DCs

Key Technologies

Network Programmability

Software Defined Networking

- Logically centralized control
- Stateless data plane (switches)
- **Edge computing:**
  - Enforce off-loading decisions: device, edge or cloud
  - Failsafe and load-balancing
  - Service migration support: restore connectivity



Data Plane Programmability

- Adding ability for local decision-making (switches)
- P4
- VNFs w/ DPDK, etc.

5GLOGINNOV

Key Technologies

Acceleration

- Virtualization and programmability come at a performance cost
  - COTS CPUs not suitable for network or AI/ML processing
- Edge computing characterized by limited processing capacity
  - Typically worse as we get closer to the user

- FPGAs *(Field Programmable Gate Arrays)*
  - Combine dedicated HW performance w/ SW flexibility
- GPUs *(Graphics Processing Units)*
  - Suitable for matrix operations (typical in AI/ML)
- TPUs *(Tensor Processing Units)* → Google Coral Edge TPU
- VPUs *(Vision Processing Units)* → Intel Neural Compute Stick

NVIDIA

NVIDIA Jetson Nano 2GB Developer Kit

intel

Intel® Neural Compute Stick 2 (Intel® NCS2)

Google®
Coral USB
Accelerator

Google Coral

Challenges

Challenges

- Extending the compute continuum
  - Making use of intermittent resources
- Mobility management in Edge computing
  - Service continuity
- Service Management and Orchestration: facing AI/ML pipelines
  - Energy
  - Accuracy
  - Privacy
- Hierarchical Inference
- Decentralization
  - Interoperability
  - Trust & Accountability

Challenges

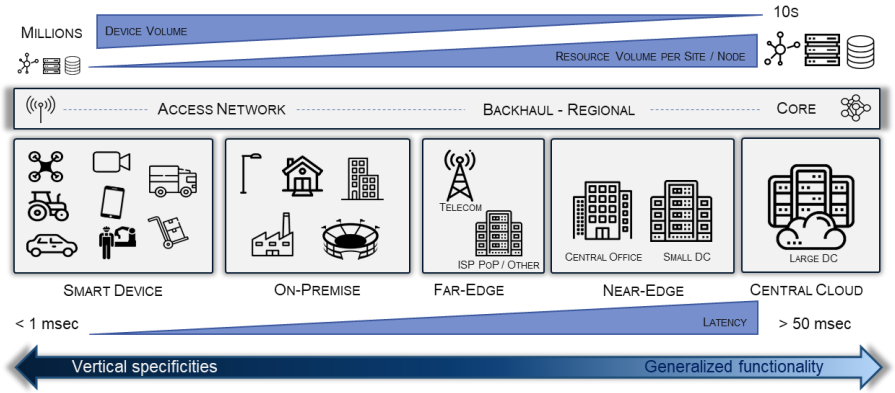
Extending the Compute Continuum

SOTA

- Lightweight virtualization and MANO e.g., Kubernetes/K3/Fog05
  - × Assumes stable connectivity, static topology

Beyond SOTA

- ✓ Asynchronous / event based MANO interfaces / APIs
  - Resource / Service discovery
  - Life-Cycle Management
- ✓ Trust/accountability mechanisms



Initial model

Resources typically provisioned by service consumers (clients)….

5GLOGINNOV

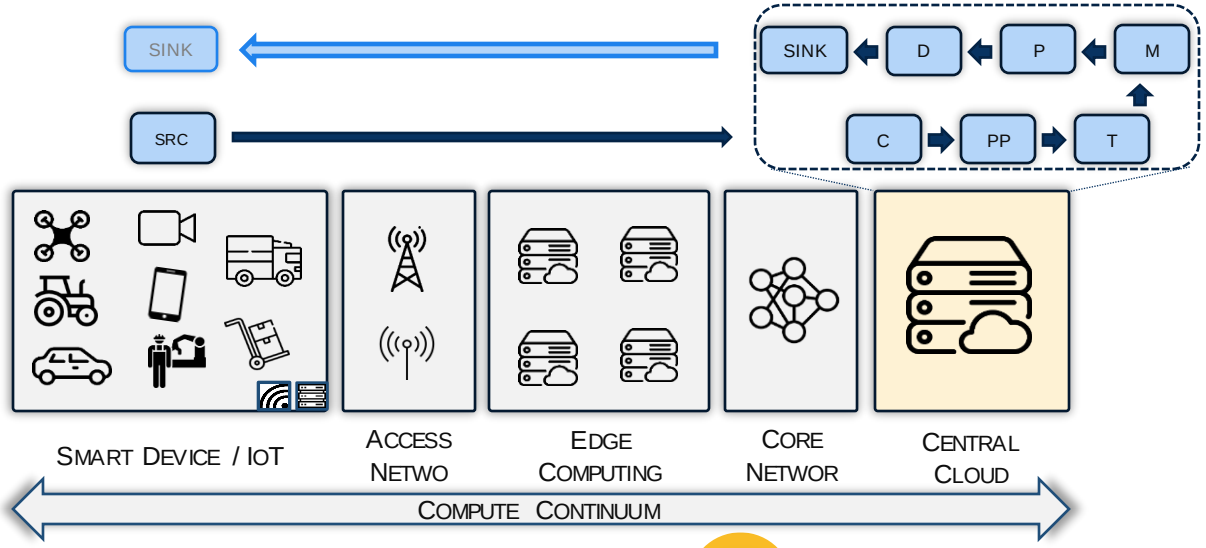
MANO for AI/ML workloads on the Compute Continuum

SoTA: AI/ML pipelines typically:
- Fully-centralized
- Static
- Reduced automation / manual
- *One-off*

- ✕ Big data overheads e.g., traffic, energy
- ✕ Limited adaptation to resource demand availability dynamics
- ✕ Privacy concerns
- ✕ Non-continuous adaptation to evolving data

- **SRC:** source of data
- **C:** collector of data from one or more sources
- **PP:** preprocessing, cleaning, feature extraction
- **T:** model training
- **M:** machine learning model
- **P:** policy / rules for safeguarding
- **D:** distributor of M output to enforcement
- **SINK:** node taking action on M output

Adapted from: Recommendation ITU-T Y.3172 (06/2019) Architectural framework for machine learning in future networks including IMT-2020



5GLOGINNOV

MANO for AI/ML workloads on the Compute Continuum

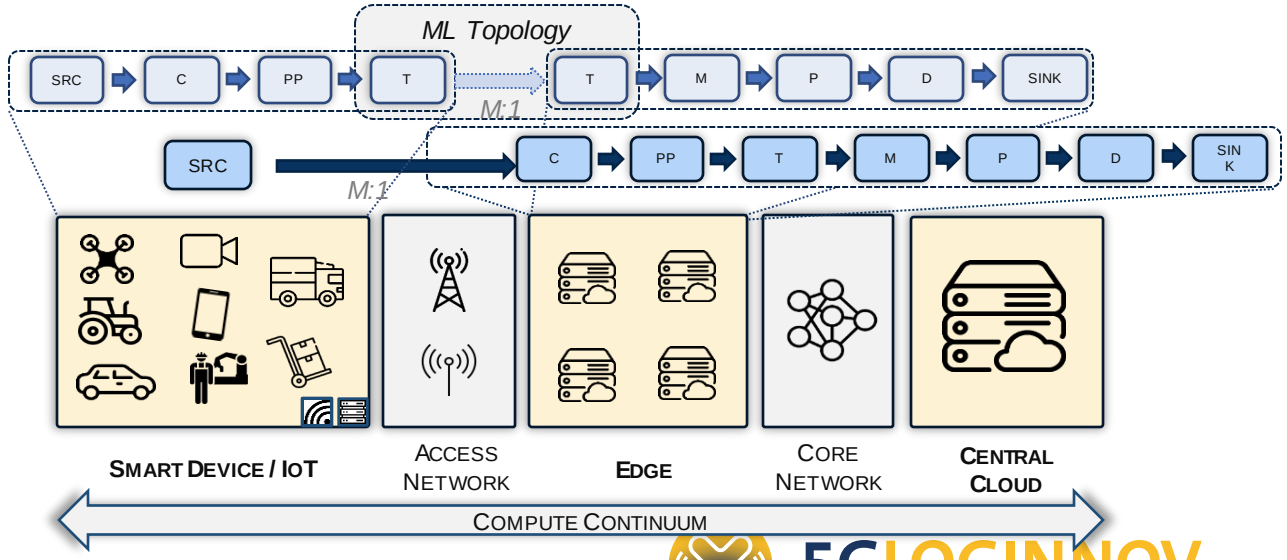
Beyond SoTA: MANO framework

- Distributed
  - Smart Device-Edge-Cloud
- Dynamic/adaptive
  - Data availability
  - Resource Availability
  - Inference demand
- Automated

- ✓ Efficient resource utilization
- ✓ Reduced latencies
- ✓ Privacy friendly
- ✓ Continuous adaptation to evolving data

- **SRC:** source of data
- **C:** collector of data from one or more sources
- **PP:** preprocessing, cleaning, feature extraction
- **T:** model training
- **M:** machine learning model
- **P:** policy / rules for safeguarding
- **D:** distributor of M output to enforcement
- **SINK:** node taking action on M output

Adapted from: Recommendation ITU-T Y.3172 (06/2019) Architectural framework for machine learning in future networks including IMT-2020



5GLOGINNOV

Thank you!

Questions?

5GLOGINNOV

Co-financed by the European Union
Connecting Europe Facility

k.katsaros@iccs.gr